D32–D37 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 11 November 2013*
doi:10.1093/nar/gkt1030

GenBank

Dennis A. Benson, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received September 18, 2013; Accepted October 7, 2013

ABSTRACT

GenBank® is a comprehensive database that contains publicly available nucleotide sequences for over 280 000 formally described species. These sequences are obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole-genome shotgun and environmental sampling projects. Most submissions are made using the web-based BankIt or standalone Sequin programs, and GenBank staff assign accession numbers upon data receipt. Daily data exchange with the European Nucleotide Archive and the DNA Data Bank of Japan ensures worldwide coverage. GenBank is accessible through the National Center for Biotechnology Information (NCBI) Entrez retrieval system, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. To access GenBank and its related retrieval and analysis services, begin at the NCBI home page: www.ncbi.nlm.nih.gov.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD, USA.

NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS), whole-genome shotgun (WGS) and other high-throughput data from sequencing centers. The US Patent and Trademark Office also contributes sequences from issued patents. GenBank participates with the EMBL Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA) (2) and the DNA Data Bank of Japan (DDBJ) (3) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC). The INSDC partners exchange data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide. NCBI makes GenBank data available at no cost over the Internet, through FTP and a wide range of Web-based retrieval and analysis services (4).

RECENT DEVELOPMENTS

Updates to the submission portal and related tools

NCBI continues to expand a unified submission portal that will ultimately provide a single access point for data submitters (submit.ncbi.nlm.nih.gov). Submitters can create accounts that track and display all their submissions and that facilitate communication with relevant NCBI staff. With respect to GenBank, the portal now supports submissions of complete microbial genomes along with WGS genomes. Submitters may continue to use standard GenBank submission tools (see below) for other GenBank submissions.

In addition to the submission portal itself, GenBank now provides some additional tools for submitters. For submitters of prokaryotic ribosomal RNA data, NCBI submission tools (see below) now perform internal checks for proper strandedness or for chimeric sequences. For submitters of bacterial genomes, NCBI provides a genome annotation pipeline described on a new prokaryotic annotation page (www.ncbi.nlm.nih.gov/genome/annotation_prok/) that also gives guidance for its use. Links to this page appear on the GenBank home page (www.ncbi.nlm.nih.gov/genbank/) and the Genome home page (www.ncbi.nlm.nih.gov/genome/).

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov

Catalog files

Beginning in release 194, GenBank discontinued the legacy 'index' files that previously accompanied the sequence data files of GenBank FTP releases. In their place, GenBank now provides a set of new 'release catalog' products that reside in a directory named 'catalog' in the FTP release. These new files are tab-delimited, and each row represents a single GenBank record and includes the accession.version, GI number, molecule type, sequence length, organism, taxonomy ID, division code and BioProject and BioSample accessions. Other sets of files associate GenBank records with gene symbols or PubMed IDs. These new files are described in more detail in the GenBank release notes (ftp.ncbi.nih.gov/genbank/gbrel.txt).

TPA assemblies

GenBank now accepts submissions of third party annotation (TPA) assemblies, which are typically assemblies constructed from existing collections of reads already in the Sequence Read Archive (SRA). Submitters are now required to supply data in binary sequence alignment/map (BAM) format that should specify the alignments of individual reads to submitted scaffolds, and these alignments may contain gaps. TPA assembly sequences will have the keyword 'TPA:assembly' and will have definition lines that begin with 'TPA_asm'. An older example of such an assembly is DAAA00000000, which contains links both to the WGS contigs (DAAA02nnnnnn) and the individual scaffolds (e.g. GK000001–GK000030).

Changes to indexing of bacterial strains

In the past, GenBank has assigned unique taxonomy IDs for each particular strain of bacteria for which there were submitted sequence data. Given the rapid increase in the diversity of bacterial strains in data being submitted, in the near future GenBank will no longer assign taxonomy IDs at the strain level. Bacterial strains that already have a unique taxonomy ID will retain them, but any submitted data for a new strain will not be given a unique taxonomy ID; rather, GenBank will assign the taxonomy ID of the species to these data. Because these new strains will also be entered in the BioSample database as unique records, BioSample will become the source of unique NCBI identifiers for individual bacteria strains and isolates.

ORGANIZATION OF THE DATABASE

GenBank divisions

GenBank assigns sequence records to various divisions based either on the source taxonomy or the sequencing strategy used to obtain the data. There are 12 taxonomic divisions [BCT, environmental sample sequence (ENV), INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL and VRT] and 6 high-throughput divisions [EST, GSS, high-throughput cDNA (HTC), high-throughput genomic (HTG) and STS]. Finally, the PAT division contains records supplied by patent offices, the transcriptome shotgun assembly (TSA) division contains sequences from TSA projects and the WGS division contains sequences from WGS projects. The size and growth of these divisions, and GenBank as a whole, are shown in Table 1.

Sequence-based taxonomy

Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (www.ncbi.nlm.nih.gov/taxonomy/) developed by NCBI in collaboration with EMBL-Bank and DDBJ and with the valuable assistance of external advisers and curators (5). Over 280 000 formally described species are represented in GenBank, and the top species in the non-WGS GenBank divisions are listed in Table 2.

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a unique identifier called an accession number that is shared across the three collaborating databases (GenBank, DDBJ and EMBL-Bank). The accession number appears on the ACCESSION line of a GenBank record and remains constant over the lifetime of the record, even when there is a change to the sequence or annotation. Changes to the sequence data itself are tracked by an integer extension of the accession number, and this *Accession.version* identifier appears on the VERSION line of the GenBank flat file. The initial version of a sequence has the extension ".1". In addition, each version of the DNA sequence is also assigned a unique NCBI identifier called a *GI* number that also appears on the VERSION line following the *Accession.version*:

```
ACCESSION AF000001
VERSION AF000001.5 GI: 7274584
```

When a change is made to a sequence in a GenBank record, a new GI number is issued to the updated sequence and the version extension of the *Accession.version* identifier is incremented. The accession number for the record as a whole remains unchanged, and will always retrieve the most recent version of the record; the older versions remain available under the old *Accession.version* identifiers and their original GI numbers. The Revision History report, available from the 'Display Settings' menu on the sequence record view, summarizes the various updates for that GenBank record.

A similar system tracks changes in the corresponding protein translations. These identifiers appear as qualifiers for CDS features in the FEATURES portion of a GenBank entry, e.g. /protein_id = "AAF14809.1". Protein sequence translations also receive their own unique GI number, which appears as a second qualifier on the CDS feature:

```
/db_xref = "GI : 6513858"
```

Citing GenBank records

Besides being the primary identifier of a GenBank sequence record, GenBank accessions are also the most

Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 197 (8/2013)	Annual increase (%)[a]
WGS	Whole-genome shotgun data	500 420 412 665	62.4
TSA	Transcriptome shotgun data	8 633 123 935	49.9
PHG	Phages	119 812 712	42.5
VRL	Viruses	1 757 202 472	22.9
BCT	Bacteria	10 281 048 518	21.8
ENV	Environmental samples	3 743 277 434	10.9
INV	Invertebrates	2 737 140 646	9.8
PAT	Patented sequences	13 290 161 247	9.7
PLN	Plants	5 963 882 822	8.8
GSS	Genome survey sequences	23 726 384 753	8.1
VRT	Other vertebrates	3 068 956 026	6.3
MAM	Other mammals	911 342 025	5.6
HTG	High-throughput genomic	25 184 819 955	3.4
HTC	High-throughput cDNA	656 196 063	2.7
UNA	Unannotated	130 510	2.1
EST	Expressed sequence tags	41 665 629 009	1.9
PRI	Primates	6 425 093 034	1.7
SYN	Synthetic	941 078 074	1.4
ROD	Rodents	4 451 315 297	0.4
STS	Sequence tagged sites	636 326 479	0.0
TOTAL	All GenBank sequences	654 613 333 676	45.1

[a]Measured relative to Release 191 (8/2012).

Table 2. Top organisms in GenBank (Release 197)

Organism	Non-WGS base pairs
Homo sapiens	17 111 514 261
Mus musculus	9 982 065 736
Rattus norvegicus	6 524 450 090
Bos taurus	5 389 432 575
Zea mays	5 071 648 554
Sus scrofa	4 889 229 566
Danio rerio	3 119 512 595
Hordeum vulgare	1 463 247 509
Strongylocentrotus purpuratus	1 435 237 072
Oryza sativa Japonica Group	1 263 872 842
Macaca mulatta	1 256 717 300
Xenopus (Silurana) tropicalis	1 249 741 450
Nicotiana tabacum	1 198 798 076
Arabidopsis thaliana	1 152 899 341
Triticum aestivum	1 139 790 400
Drosophila melanogaster	1 127 199 957
Vitis vinifera	1 069 944 084
Pan troglodytes	1 008 818 677
Solanum lycopersicum	966 234 744
Canis lupus familiaris	952 526 510

efficient and reliable way to cite a sequence record in publications. We certainly encourage submitters and other authors to cite GenBank data using these accessions. However, as discussed earlier, since searching with a GenBank accession number will retrieve the most recent version of a record, the data returned from such searches will change over time if the record is updated. It is quite possible, therefore, for the sequence data retrieved today by an accession to be different from that discussed or analyzed in a paper published several years ago. We therefore recommend that authors include the version suffix when citing a GenBank accession (e.g. AF000001.5), so that future readers can easily retrieve the data in question.

BUILDING THE DATABASE

The data in GenBank and the collaborating databases, EMBL-Bank and DDBJ, are submitted either by individual authors to one of the three databases or by sequencing centers as batches of EST, STS, GSS, HTC, TSA, WGS or HTG sequences. Data are exchanged daily with DDBJ and EMBL-Bank so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct electronic submission

Virtually all records enter GenBank as direct electronic submissions (www.ncbi.nlm.nih.gov/genbank/), with the majority of authors using the BankIt or Sequin programs. Many journals require authors with sequence data to submit the data to a public sequence database as a condition of publication. GenBank staff can usually assign an accession number to a sequence submission within two working days of receipt, and do so at a rate of ~3500 per day. The accession number serves as confirmation that the sequence has been submitted and provides a means for readers of articles in which the sequence is cited to retrieve the data. Direct submissions receive a quality assurance review that includes checks for vector contamination, proper translation of coding regions, correct taxonomy and correct bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database.

Authors may ask that their sequences be kept confidential until the time of publication. Since GenBank policy requires that the deposited sequence data be made public when the sequence or accession number is published, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited to ensure a timely release of the data. Although only the submitter is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank at update@ncbi.nlm.nih.gov.

NCBI works closely with sequencing centers to ensure timely incorporation of bulk data into GenBank for public release. GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program *tbl2asn*, described at www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html. Submitters can keep abreast of updates to *tbl2asn* and Sequin by subscribing to the NCBI submissions RSS feed (www.ncbi.nlm.nih.gov/feed/rss.cgi?ChanKey=genbanksubmissiontoo).

Submission using BankIt

About a third of author submissions are received through an NCBI Web-based data submission tool named BankIt. Using BankIt, authors enter sequence information and biological annotations, such as coding regions or mRNA features, directly into a series of tabbed forms that allow the submitter to describe the sequence further without

having to learn formatting rules or controlled vocabularies. In addition, BankIt allows submitters to upload source and annotation data using tab-delimited tables. Before creating a draft record in the GenBank flat file format for the submitter to review, BankIt validates the submissions by flagging many common errors and checking for vector contamination using a variant of BLAST called Vecscreen.

Submission using Sequin and tbl2asn

NCBI also offers a standalone multi-platform submission program called Sequin (www.ncbi.nlm.nih.gov/projects/Sequin/) that can be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences (such as a single cDNA), phylogenetic studies, population studies, mutation studies, environmental samples with or without alignments and sequences with complex annotation. Sequin has convenient editing and complex annotation capabilities and contains a number of built-in validation functions for quality assurance. In addition, Sequin is able to accommodate large sequences, such as the 5.6-Mb *Escherichia coli* genome, and read in a full complement of annotations from simple tables. The most recent version, Sequin 12.30, was released in November 2012 and is available for Macintosh, PC and Unix computers via anonymous FTP at ftp.ncbi.nlm.nih.gov/sequin. Once a submission is completed, submitters can e-mail the Sequin file to gb-sub@ncbi.nlm.nih.gov or upload the Sequin file to www.ncbi.nlm.nih.gov/LargeDirSubs/dir_submit.cgi. Submitters of large, heavily annotated genomes may find it convenient to use the command line tool *tbl2asn* to convert a table of annotations generated from an annotation pipeline into an ASN.1 (Abstract Syntax Notation One) record suitable for submission to GenBank. These files for WGS genome and TSA submissions are then transmitted to GenBank via the Submission Portal.

Notes on particular divisions

Environmental sample sequences

The ENV division of GenBank accommodates sequences obtained via environmental sampling methods in which the source organism is unknown. Many ENV sequences arise from metagenome samples derived from microbiota in various animal tissues, such as within the gut or skin, or from particular environments, such as freshwater sediment, hot springs or areas of mine drainage. Records in the ENV division contain 'ENV' in the keyword field and use an "/environmental_sample" qualifier in the source feature. Environmental sample sequences are generally submitted for whole metagenomic shotgun sequencing experiments or surveys of sequences from targeted genes, like 16S rRNA. NCBI continues to support BLAST searches (see below) of metagenomic ENV sequences, but sequences within WGS projects are now part of the WGS BLAST database.

Whole-genome shotgun sequences

WGS sequences appear in GenBank as groups of sequence-overlap contigs collected under a master WGS record. Each master record represents a WGS project and has an accession number in the Nucleotide database consisting of a four-letter prefix followed by eight zeroes and a version suffix as found in standard GenBank records. The number of zeroes increases to nine for WGS projects with 1 million or more contigs. Master records contain no sequence data; rather, links appear at the bottom of these records that provide displays of individual contigs in the WGS browser. Contig records have accessions consisting of the same four-letter prefix as their master accession, followed by a two-digit version number and a six-digit contig ID. For example, the WGS accession number 'AAAA02002744' is assigned to contig number '002744' of the second version of project 'AAAA', whose accession number is 'AAAA00000000.2'. Currently, there are over 12 000 WGS sequencing projects, many of whose data have been used to build over 14 million scaffolds and chromosomes for genome assemblies. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/Traces/wgs/.

Although WGS project sequences may be annotated, many low-coverage genome projects do not contain annotation. Because these sequence projects are ongoing and incomplete, these annotations may not be tracked from one assembly version to the next and should be considered preliminary. Submitters of genomic sequences, including WGS sequences, are urged to use evidence tags of the form "/experimental = *text*" and "/inference = *TYPE*: *text*", where *TYPE* is one of a number of standard inference types and *text* consists of structured text.

Transcriptome shotgun assembly sequences

The TSA division contains TSA sequences that are assembled from sequences deposited in the NCBI Trace Archive, the SRA and the EST division of GenBank. While neither the Trace Archive nor SRA is a part of GenBank, they are part of the INSDC and provide access to the data underlying these assemblies (4,6). TSA records have 'TSA' as their keyword and can be retrieved with the query 'tsa[properties]'. TSA continues to be one of the most rapidly growing divisions of GenBank, growing by 50% in the past year.

Genome survey sequences

GSS data continue to be an important resource for genome sequencing and annotation studies, and at almost 24 billion base pairs, the GSS division remains one of the largest non-WGS divisions in GenBank. GSS data are available for download from ftp.ncbi.nlm.nih.gov/repository/dbGSS/ and from the GenBank FTP site. In the near future GenBank will begin accepting GSS sequences with annotations provided by submitters. Because the data model for dbGSS does not allow for annotations, these new sequences will be indexed in the Entrez Nucleotide database (see below) and can be retrieved with the query 'gbdiv gss[prop]'.

High-throughput genomic and high-throughput cDNA sequences

The HTG division of GenBank (www.ncbi.nlm.nih.gov/genbank/htgs/) contains unfinished large-scale genomic

records, which are in transition to a finished state (7). These records are designated as belonging to Phases 0–3 depending on the quality of the data, with Phase 3 being the finished state. Upon reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank.

The HTC division of GenBank contains HTC sequences that are of draft quality but may contain 5′ UTRs, 3′ UTRs, partial coding regions and introns. HTC sequences which are finished and of high quality are moved to the appropriate organism division of GenBank. A project generating HTC data is described in (8).

Special Record types

Third party annotation

TPA records are sequence annotations published by someone other than the original submitter of the primary sequence record in DDBJ/ENA/GenBank (www.ncbi.nlm.nih.gov/genbank/TPA). In addition to the new TPA assembly sequences discussed earlier, each of the 167 000 TPA records falls into one of three categories: 'experimental', in which case there is direct experimental evidence for the existence of the annotated molecule; 'inferential', in which case the experimental evidence is indirect; and 'assembly', where the focus is on providing a better assembly of the raw reads. TPA sequences may be created by assembling a number of primary sequences. The format of a TPA record (e.g. BK000016) is similar to that of a conventional GenBank record but includes the label "TPA_exp:", "TPA_inf:" or "TPA_asm:" at the beginning of each Definition Line as well as corresponding keywords. TPA experimental and inferential records also contain a Primary block that provides the base ranges and identifier for the sequences used to build the TPA. TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. TPA submissions to GenBank may be made using either BankIt or Sequin.

Contig (CON) records for assemblies of smaller records

Within GenBank, CON records are used to represent very long sequences, such as a eukaryotic chromosome, where the sequence is not complete but consists of several contig records with uncharacterized gaps between them. Rather than listing the sequence itself, CON records contain assembly instructions involving the several component sequences. An example of such a CON record is CM000663 for human chromosome 1.

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible through the NCBI Entrez retrieval system (4). Records from the EST and GSS divisions of GenBank are stored in the EST and GSS databases, whereas all other GenBank records are stored in the Nucleotide database. GenBank sequences that are part of population or phylogenetic studies are also collected together in the PopSet database, and conceptual translations of CDS sequences annotated on GenBank records are available in the Protein database. Each of these databases is linked to the scientific literature in PubMed and PubMed Central. Additional information about conducting Entrez searches is found in the NCBI Help Manual (www.ncbi.nlm.nih.gov/books/NBK3831/) and links to related tutorials are provided on the NCBI Education page (www.ncbi.nlm.nih.gov/education/).

Associating sequence records with sequencing projects

The ability to identify all GenBank records submitted by a specific group or those with a particular focus, such as metagenomic surveys, is essential for the analysis of large volumes of sequence data. The use of organism or submitter names as a means to define such a set of sequences is unreliable. The BioProject database (www.ncbi.nlm.nih.gov/bioproject), developed at NCBI and subsequently adopted across the INSDC, allows submitters to register large-scale sequencing projects under a unique project identifier, enabling reliable linkage between sequencing projects and the data they produce. BioProject includes pointers to data from a wide variety of projects deposited in any NCBI primary data archive. Sequencing projects focus on genomes, metagenomes, transcriptomes, comparative genomics and on particular loci, such as 16S ribosomal RNA. A 'DBLINK' line appearing in GenBank flat files identifies the sequencing projects with which a GenBank sequence record is associated. In addition, sequence records may now have a link to the BioSample database (9) that provides additional information about the biological materials used in the study that produced the sequence data. Such studies include genome wide association studies, high-throughput sequencing, microarrays and epigenomic analyses. As an example, the TSA project GAAA contains DBLINK lines that associate the GenBank sequence record with BioProject record PRJNA54005 and BioSample record SRS283232, as well as the SRA record containing the raw data, SRR401852:

```
BioProject: PRJNA77699
BioSample: SRS283232
Sequence Read Archive: SRR401852
```

BLAST sequence-similarity searching

Sequence-similarity searches are the most fundamental and frequent type of analysis performed on GenBank data. NCBI offers the BLAST family of programs (blast.ncbi.nlm.nih.gov) to detect similarities between a query sequence and database sequences (10,11). BLAST searches may be performed on the NCBI Web site (12) or by using a set of standalone programs distributed by FTP (4).

Obtaining GenBank by FTP

NCBI distributes GenBank releases in the traditional flat file format as well as in the ASN.1 format used for internal maintenance. The full bimonthly GenBank release along with the daily updates, which incorporate sequence data

from EMBL-Bank and DDBJ, is available by anonymous FTP from NCBI at ftp.ncbi.nlm.nih.gov/genbank. The full release in flat file format is available as a set of compressed files with a non-cumulative set of updates at ftp.ncbi.nlm.nih.gov/genbank/daily-nc/. For convenience in file transfer, the data are partitioned into multiple files; for release 197 there are 1905 files requiring 607 GB of uncompressed disk storage. A script is provided in ftp.ncbi.nlm.nih.gov/genbank/tools/ to convert a set of daily updates into a cumulative update.

MAILING ADDRESS

GenBank, National Center for Biotechnology Information, Building 45, Room 6AN12D-37, 45 Center Drive, Bethesda, MD 20892, USA.

ELECTRONIC ADDRESSES

www.ncbi.nlm.nih.gov: NCBI Home Page.
gb-sub@ncbi.nlm.nih.gov: Submission of sequence data to GenBank.
update@ncbi.nlm.nih.gov: Revisions to, or notification of release of, 'confidential' GenBank entries.
info@ncbi.nlm.nih.gov: General information about NCBI resources.

CITING GENBANK

If you use the GenBank database in your published research, we ask that this article be cited.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
2. Cochrane,G., Alako,B., Amid,C., Bower,L., Cerdeno-Tarraga,A., Cleland,I., Gibson,R., Goodgame,N., Jang,M., Kay,S. *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.*, **41**, D30–D35.
3. Ogasawara,O., Mashima,J., Kodama,Y., Kaminuma,E., Nakamura,Y., Okubo,K. and Takagi,T. (2013) DDBJ new system and service refactoring. *Nucleic Acids Res.*, **41**, D25–D29.
4. NCBI Resource Coordinators. (2013) Database resources at the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
5. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
6. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
7. Kans,J.A. and Ouellette,B.F.F. (2001) Submitting DNA sequences to the databases. In: Baxevanis,A.D. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc, New York, NY, pp. 65–81.
8. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
9. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
12. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.